PROSODIC AIDS TO SPEECH RECOGNITION. II. SYNTACTIC SEGMENTATION AND STRESSED SYLLABLE LOCATION

Wayne A. Lea, et al

Sperry Rand Corporation

Prepared for:

Advanced Research Projects Agency

15 April 1973

SPERRY UNIVAC

PROSODIC AIDS TO SPEECH RECOGNITION:

II. SYNTACTIC SEGMENTATION AND STRESSED SYLLABLE LOCATION

by

Wayne A. Lea
Mark F. Medress
Toby E. Skinner

Defense Systems Division
St. Paul, Minnesota
(612-456-2430)

Final Technical Report Submitted To:

Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

Attention: Dr. L. G. Roberts

15 April 1973

Report No. PX 10232

This research was supported by the Advanced Research Projects Agency of the Department of Defense under Contract No. DAHC 15-72-C-0138, ARPA Order No. 2010, Program Code No. 90536. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U. S. Government.

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

Univac Defense Systems Division
P.O. Box 3525
St. Paul, Minnesota 55165

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

Prosodic Aids to Speech Recognition II: Syntactic Segmentation and Stressed Syllable Location

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Final Technical Report; 1 March 1972 – 28 February 1973

5. AUTHOR(S) *(First name, middle initial, last name)*

1) Wayne A. Lea
2) Mark F. Medress
3) Toby E. Skinner

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
15 April 1973	34	18

8a. CONTRACT OR GRANT NO	9a. ORIGINATOR'S REPORT NUMBER(S)
DAHC15-72-C-0138	Univac Report No. PX 10232
b. PROJECT NO	
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	None

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, Virginia 22209

13. ABSTRACT

A strategy is outlined for acoustic aspects of speech recognition, whereby prosodic features are used to detect boundaries between phrases, then stressed syllables are located within each constituent and a partial distinctive feature analysis is done within stressed syllables. Facilities have been implemented for linear prediction, formant tracking, and extraction of fundamental frequency and speech energy contours. Experiments were conducted on the automatic detection of constituent boundaries and location of stressed syllables by analysis of fundamental frequency and energy contours, for recordings of six talkers reading the Rainbow Script, two talkers reading a paragraph composed of monosyllabic words, and ten talkers involved in speaking sentences pertinent to man-computer interaction. A program was implemented which successfully detects over 80% of all boundaries between major syntactic constituents, by the use of fall-rise valleys in fundamental frequency contours. A panel of three listeners provided judgments of which syllables were stressed, unstressed, or reduced in the speech texts. Judgments from two listeners were quite consistent from time to time, and the two listeners particularly agreed with each other as to which syllables were stressed. The third listener gave less consistent results. An algorithm was devised for locating stressed syllables as high energy portions of speech with rising or non-falling fundamental frequency. This algorithm succeeded in locating 85% of all syllables that had been perceived as stressed by two or more listeners. Further work will involve implementation of the stressed syllable location algorithm, refinements of syntactic boundary predictions and detection procedures, further tests with designed speech texts, and applications to distinctive features estimation and syntactic parsing.

DD FORM 1 NOV 65 1473

Unclassified
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech Perception						
Speech Recognition						
Speech Analysis						
Linguistic Stress						
Prosodies						
Prosodic Features Extraction						
Syntactic Boundary Detection						
Distinctive Features Estimation						
Syntactic Analysis						
Syntactic Parsing						

SPERRY UNIVAC

PROSODIC AIDS TO SPEECH RECOGNITION:

II. SYNTACTIC SEGMENTATION AND STRESSED SYLLABLE LOCATION

by

Wayne A. Lea
Mark F. Medress
Toby E. Skinner

Defense Systems Division
St. Paul, Minnesota
(612-456-2430)

Final Technical Report Submitted To:

Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

Attention: Dr. L. G. Roberts

15 April 1973

Report No. PX 10232

This research was supported by the Advanced Research Projects Agency of the Department of Defense under Contract No. DAHC 15-72-C-0138, ARPA Order No. 2010, Program Code No. 90536. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U. S. Government.

PREFACE

This is the second in a series of reports on Prosodic Aids to Speech Recognition. The first report, subtitled, "I. Basic Algorithms and Stress Studies," appeared 1 October 1972, as Univac Report No. PX 7940. (The subtitle did not appear on all copies of that report.)

This research was supported by the Advanced Research Projects Agency of the Department of Defense, under Contract No. DAHC15-72-C-0138, ARPA Order No. 2010, Program Code No. 90536. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Advanced Research Projects Agency or the U. S. Government.

SUMMARY

Computers that understand speech are expected to facilitate natural man-machine interaction, but the problems involved demand the attention of several disciplines, including linguistics, computer systems design, perception theory, speech research, and engineering. Linguistic and perceptual arguments, in particular, suggest that devices which recognize speech will have to make use of grammatical structure ("syntax") in early stages of the recognition procedures (Lea, 1972a,b; 1973b; Lea, Medress, and Skinner, 1972a). This can be accomplished, in part, by using certain acoustic features (called "prosodic features") to segment the speech into grammatical phrases, and to identify those syllables that are given prominence, or stress, in the sentence structure.

Once the connected speech is segmented into phrases, and stressed syllables are located, the Univac speech recognition strategy would call for a partial distinctive features analysis within each stressed syllable. Consonants and vowels are expected to be more clearly articulated and easier to distinguish in stressed syllables, than in unstressed or reduced syllables (cf. Lea, Medress, and Skinner, 1972b), where articulation (and consequent acoustic information) is not as precise or consistent from talker to talker or time to time.

An algorithm has been devised for segmenting speech into grammatical phrases, by marking phrase boundaries at the bottoms of fall-rise valleys in fundamental frequency (F_0) contours. This algorithm has been implemented as a FORTRAN program on the Univac interactive speech research facility. It uses F_0 data obtained from a Univac fundamental frequency tracking program which works on the principle of autocorrelation of the speech waveform (see Appendix A of this report).

A strategy for locating stressed syllables has also been devised. Based on previous studies that have shown that local increases in F_0 and large integrals of energy within a syllable are the most reliable acoustic correlates of stress, this algorithm looks for regions of high energy integral near local F_0 increases. It also incorporates adjustments based on the most common ("archetype") F_0 contours within the grammatical phrases and clauses of connected speech. This strategy has been specified precisely, but has not yet been implemented as a computer program. It uses the F_0 data from the fundamental frequency tracker and an (essentially broad-band) energy measure obtained from the speech waveform.

To test these algorithms with actual speech, several spoken texts were selected. The "Rainbow Script", a short six-sentence paragraph used extensively in speech studies, was recorded by six talkers. A paragraph composed of only monosyllabic words was also recorded, by two talkers. Finally, thirteen sentences were selected from recordings by five ARPA contractors. These ARPA sentences involve various questions, commands, and declarations pertinent to man-machine interaction, and most were taken from simulated or actual man-computer protocols.

The program for detecting syntactic boundaries from fall-rise patterns in voice fundamental frequency contours has been shown, both by the present study with these spoken texts and by previous studies, to succeed in finding about 80% of all syntactically predicted boundaries between major syntactic units. It also, however, detects some syntactic boundaries not predicted by an intuitive constituent structure analysis, and detects a few false boundaries not apparently related to syntactic structure, such as at consonant-vowel boundaries.

The algorithm for stressed syllable location succeeded in locating around 85% of all syllables perceived as stressed by the majority votes of a panel of listeners.

Perceptions of stress levels by three listeners proved to be an adequate standard for testing the algorithmic location of stressed syllables. Two listeners were found to agree in their perceived stress levels for most of the individual syllables in the Rainbow Script and Monosyllabic Script, and ARPA man-machine interaction sentences. They differed on only about 5% of all syllables as to whether they were stressed or not, and each of them showed only about 5% confusions in decisions about stressed syllables from one trial to another. Unstressed and reduced levels were much more frequently confused. A third listener differed from the other two listeners on about half of his stress level judgments, and also labelled substantial percentages of all syllables as stressed on one trial and unstressed on another. Such listeners who are inconsistent in their own judgments and who differ dramatically from other listeners should be excluded in any attempts to establish standards about which are the actual "stressed syllables" in connected speech.

The listeners appear to be as consistent in their assignments of stress levels given only the written text as they are in their assignments when listening to the speech recordings. However, their judgments without speech do not correspond well with their

judgments with speech if the speech is spontaneous (that is, not produced by speakers reading written texts). Listeners apparently differ most dramatically from each other, and yield more confusion in stress levels from repetition to repetition, when yes-no questions are involved.

Further work is needed to improve the boundary detector and to implement the stressed syllable algorithm. Linguistic predictions of boundaries and stressed syllables must yet be developed. Controlled experiments must be done on boundary detection and stressed syllable location in extensive texts which are explicitly designed to isolate effects of various sentence types, phrase structures, words, and sound sequences. Techniques must be developed using syntactic segmentation and stressed syllable location to aid partial distinctive features estimation and syntactic parsing.

TABLE OF CONTENTS

1. INTRODUCTION

This is a report on work currently in progress in the Univac Speech Communications Group, under contract with the Advanced Research Projects Agency (ARPA). As a part of ARPA's total program in research on speech understanding systems, the research reported herein is concerned with extracting reliable prosodic and distinctive features information from the acoustic waveform of connected speech (sentences and discourses). Studies are being concentrated on problems of detecting stressed syllables and syntactic boundaries.

At Univac, the viewpoint is that versatile speech recognition will proceed by making use of reliable information in the acoustic data, in combination with early use of linguistic regularities. As has been outlined in a previous report (Lea, Medress, and Skinner, 1972a), recognition is to be accomplished by using prosodically-detected stress patterns and syntactic structure in aiding a partial distinctive-feature-estimation procedure. Prosodically-detected syntactic structure will also be used to aid syntactic parsers and semantic processors.

Prosodic cues to sentence structure, and prosodic aids to the location of reliable acoustic phonetic information, have been given little or no attention in previous speech recognition efforts. The strong motivations for the use of prosodic patterns in speech recognition procedures was thus presented in some detail in the earlier report (Lea, Medress, and Skinner, 1972a, section 2). Recent improvements in the Univac facilities for extracting prosodic features, spectral data, and formants, and a program for detecting boundaries between syntactic phrases (constituents), will be described in section 2 of this report. Extensive experiments are described in section 3, which were conducted to: (1) determine the success of detecting boundaries between major syntactic units from fall-rise patterns in fundamental frequency contours; (2) determine listeners' abilities to perceive stressed, unstressed, and reduced syllables in read texts and spontaneous utterances; and (3) determine the success of locating stressed syllables by an algorithm which uses rising fundamental frequency and high energy integral as major acoustic correlates of stressed syllables in the constituents delimited by the boundary detector. The conclusions from the experiments and the work on feature extraction systems are summarized in section 4, and considerable further work is

suggested there. An appendix is included which summarizes the Univac method for fundamental frequency tracking, which has proven to be very accurate and dependable, and which has recently been adopted by several other ARPA contractors.

2. SYSTEMS FOR EXTRACTING PROSODIC AND DISTINCTIVE FEATURES

Basic research tools have been developed for providing a versatile capability to extract important features from continuous speech. These tools include a total interactive speech research facility and specific programs, implemented on that facility, to provide linear prediction and formant tracking, extraction of prosodic features, and detection of syntactic boundaries.

2.1 Interactive Speech Research Facility

The Univac speech research facility has been described in an earlier ARPA report (Lea, Medress, and Skinner, 1972a), and is shown again in Figure 1. The facility provides for the digitization, analysis, and subsequent display of time and spectral waveforms. Interactive control is provided by toggle switches, pushbuttons, potentiometers, and an alphanumeric display and keyboard. Analysis capabilities include Fast Fourier Transforms (FFT), linear prediction, formant tracking, fundamental frequency detection, energy computation, and syntactic boundary detection. The program for stressed syllable location will also be implemented on the research facility.

A Very Distant Host connection to the ARPANET is being designed and implemented at Univac, under Univac funding. An available Univac 1218 computer will serve much like the usual Terminal Interface Message Processor (TIP), but will not have packet forwarding and routing responsibilities, since it is the end of a Very Distant Host circuit.

The necessary interface hardware, which has as its main function the handling of the cyclic redundancy check and the transparent transmission conventions, has been designed, wire-wrapped, and partially debugged. The modem has been received and installed. The software for the 1218, which consists of a network control program and local terminal handlers, has been designed and programmed. Debugging of the software is now underway.

After the ARPANET interface hardware has been completed and debugged, live tests will be performed over the communications circuit to Champaigne-Urbana, Illinois. Later, local ports into the 1218 will be added, such as a modem allowing

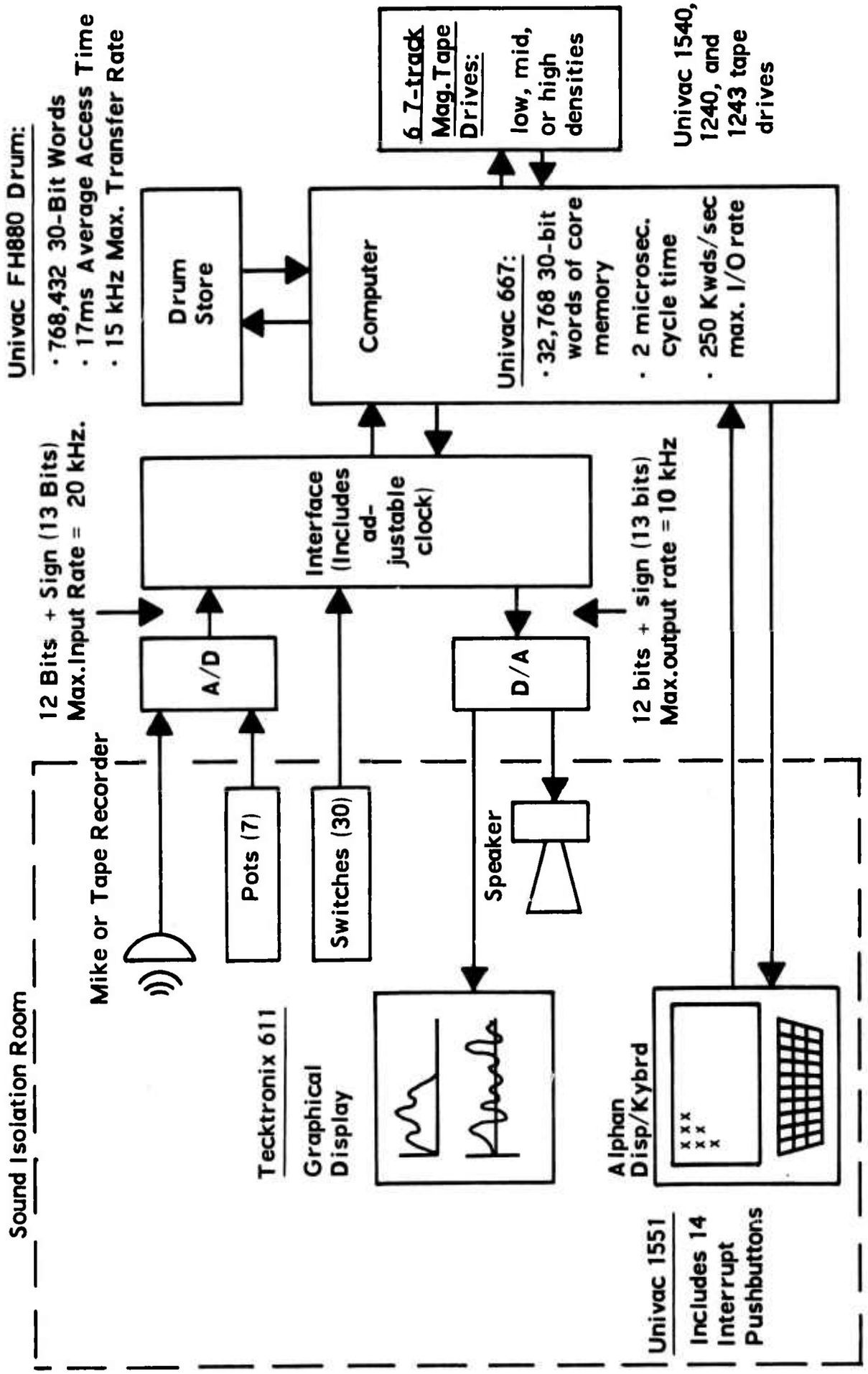

Figure 1. Block Diagram of the Univac Interactive Speech Research Facility.

any local dial-up terminal to access the 1218. Additional software will be written to allow such higher-level protocols as the File Transfer Protocol. Connections to other local computers are also being considered.

This ARPANET connection will permit access to the Lincoln Laboratories speech data base and the other contractors' programs and hardware for speech understanding research.

2.2 Linear Prediction and Formant Tracking

Critical aspects of the articulation of speech are known to be conveyed by the frequency spectrum of the speech. In particular, the formant structure found in frequency spectra conveys information about vocal tract resonances, which are expected to be useful in distinctive features analysis.

The smoothed frequency spectra needed for formant tracking are obtained from linear prediction analysis, implemented on the speech research facility. The current implementation of the spectral computation process consists of the following steps:

- low-pass filtering at 4782 Hz (using a seventh order elliptic function Cauer low-pass filter provided by Lincoln Laboratories);
- sampling at a rate of ten thousand samples per second (thus yielding a 5KHz frequency analyzing bandwidth);
- software pre-emphasis by first order differencing;
- applying a Hanning weighted analyzing time window of width 25.6 milliseconds (ms) with a 10 ms advance; and
- performing a 256 point Fast Fourier Transform.

Fourteen predictor coefficients and evaluation of the spectrum at -75 Hz off the jw-axis are used in the linear prediction procedure (Makhoul and Wolf, 1972).

In light of the successful formant tracking results obtained by other ARPA researchers who have employed spectral peak-picking techniques, the Schafer and Rabiner algorithm (Lea, Medress, and Skinner, 1972a p. 22), with its complex heuristics, has been discarded. Instead, simple peak-picking is being used for formant estimation. Recent results by other ARPA researchers have also demonstrated the value of software pre-emphasis, evaluation of the frequency spectrum off the jw-axis (Makhoul and Wolf, 1972), and global time smoothing. Pre-emphasis has been shown to aid in resolving

high-frequency formants and eliminating low frequency spectral peaks which might otherwise be confused with the first formant. Off-axis evaluation yields a spectrum with formants of narrower bandwidth and thus greater resolution. Time smoothing corrects those formant values which may be misaligned due to the simple association of peaks with formant values (FORMANT (l) = PEAK (l); l = 1, 2, 3). These improvements to formant estimation have been implemented on the Univac speech research facility.

Figure 2 shows the results of the formant tracking procedure for the utterance, "Who is the owner of utterance eight?" (ARPA test utterance LS21). Each 10 ms during voiced intervals (as defined by the fundamental frequency determination procedure), frequency values in Hertz for the first three formants (F_1, F_2, F_3) are graphed and tabulated. Also tabulated, in synchrony with the formant frequencies, is the fundamental frequency (F_0) in Hertz. For example, at time 500 ms, F_0 =107 Hz, F_1 = 429 Hz, F_2 = 1757 Hz and F_3 = 2304 Hz. It can be seen that formants close in frequency are resolved, in the vicinities of 470, 760, 1100 and 1540 ms. Also, formant values depict well the phonemic content of the utterance (e.g. /i/ at 460 ms, /n/ at 720 ms). However, some problems still exist in the formant tracking procedure as evidenced by the algorithm's confusion of the fourth formant with F_3 at times 580 to 600 ms and 1610 to 1620 ms.

In summary, substantial progress in formant tracking has been made at Univac during the past six months. The present formant tracking procedure with these refinements appears to perform well, as demonstrated by results on the ARPA test sentences. Further refinements will be incorporated as needed.

2.3 Prosodic Features Extraction

Energy and fundamental frequency time functions are computed for prosodic features analysis. Total energy is computed every 10 ms from the sum of the squares of time samples in a 25.6 ms Hanning weighted time waveform, which has not been pre-emphasized. Fundamental frequency (F_0) is computed by autocorrelating the center-clipped acoustic time waveform every 10 ms for a 51.2 ms rectangular time window over a range of 60 to 400 Hz. (A formal description of the procedure is given in the Appendix.) Although the range limits on F_0 are variable inputs to the analysis procedure, both computation time and the potential for octave errors are lessened

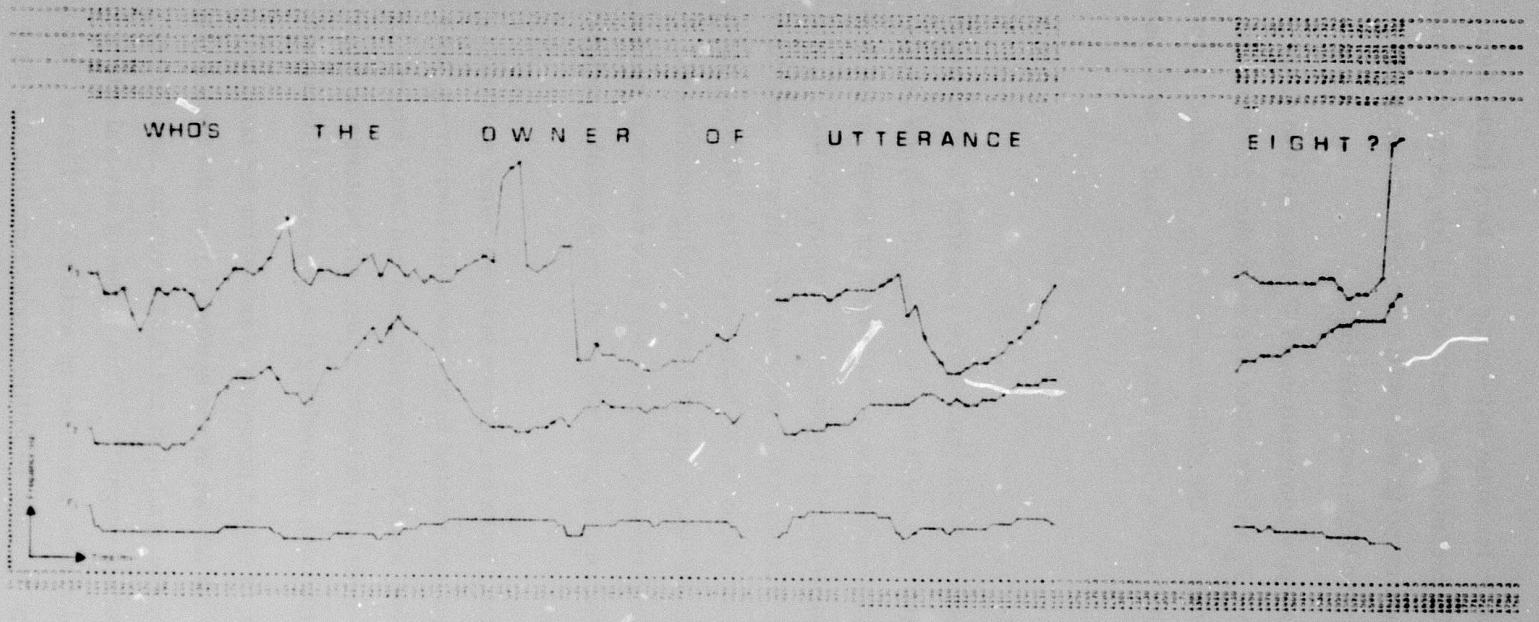

Figure 2. The First Three Formant Contours for the Utterance "Who is the Owner of Utterance Eight?" (LS21)

Tabulated at the top of the graphs are F_0 (bottom-most, or closest to the graphs), and F_1, F_2, and F_3 (in ascending order), all in Hertz.

by using separate ranges of F_0 search for male and female speakers (male: 60 to 230 Hz; female 140 to 400 Hz). This technique for fundamental frequency determination has been experimentally demonstrated to work well even when the analyzed utterance has been subjected to severe high-pass filtering.

Following independent time window computations of F_0, one and two value global time smoothing is applied to correct F_0 values which are more than 20 Hertz misaligned from their neighboring values. Of the 40308 values of F_0 computed independently for the various texts, only 305 values (less than 1%) were changed as a result of the smoothing process.

2.4 Syntactic Boundary Detection

The prosodic patterns obtained from the interactive research facility are used as inputs to Lea's program for detecting boundaries between major syntactic constituents (Lea,1972b,Chapter 2), implemented as a FORTRAN program on the research facility. The boundary-detection algorithm is based on an assumption that F_0 will usually decrease (about 7% or more) at the end of each major syntactic constituent, and then increase (about 7% or more) either at the beginning of the following constituent or after any unstressed syllables at the beginning of that following constituent. Experimenting with fundamental frequency contours in over 500 seconds of speech, Lea (1972b,1973a) had previously shown that over 80% of all syntactically predicted boundaries were correctly detected.

In these previous studies, some extra boundaries between minor syntactic constituents and some false boundaries (not related to any syntactic boundaries, but rather resulting from F_0 variations between vowels and consonants) had been detected by the algorithm. The algorithm also successfully detected clause and sentence boundaries wherever long (350 millisecond) stretches of unvoicing (i.e., "pauses") occurred.

Figure 3 illustrates the form of output presently obtained for the boundary detection program. The total energy in dB (shown by the symbol +) and the fundamental frequency in eighth tones (shown by the symbol 0) are plotted versus time for the sentence "Who is the owner of utterance eight?". This is one of several utterances recorded by ARPA contractors, as will be described in section 3.1. Figure 3 shows a value recorded for F_0 and energy for each 10 milliseconds from time 0 to time 1760 milliseconds. The tabular data at the top of the graph are broadband energy in dB, next, fundamental frequency in Hertz, and finally (uppermost of the three),

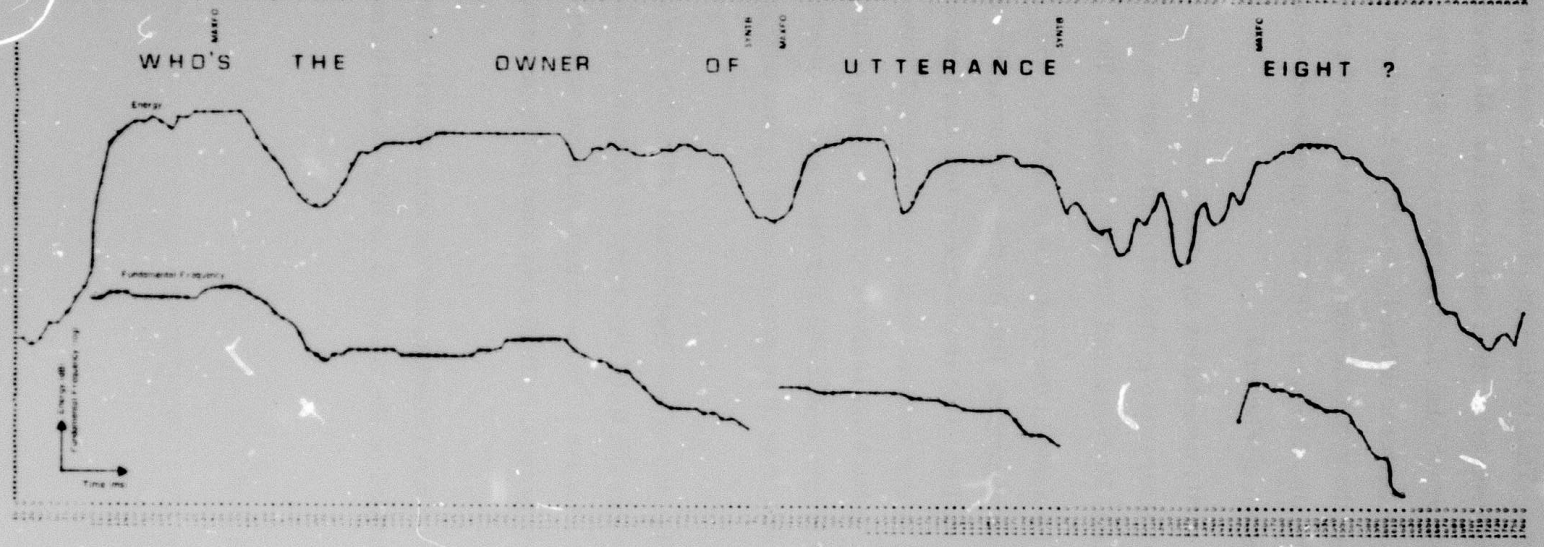

Figure 3. Fundamental Frequency (Symbolized by 0's) and Energy (Symbolized by +'s) Versus Time for the Utterance "Who is the Owner of Utterance Eight?" (LS21)

Tabulated at the top of the graphs are fundamental frequency in eighth tones (uppermost), fundamental frequency in Hertz (next down), and energy in dB (nearest the graphs). Changes in F_0 of about 5 eighth tones correspond to the 7% variations needed for constituent boundary detection. Detected syntactic boundaries are marked by SYNTB and the (first occurence of) maximum F_0 within each constituent is marked by MAXFO.

fundamental frequency in eighth tones. This is the type of data that has previously been available from the prosodic features extraction facilities (Lea, Medress, and Skinner, 1972a, p. 25). However, the graph also displays the mark "SYNTB" at each position where the syntactic boundary detector located a syntactic boundary, and the mark "MAXFO" at the first point in the syntactic constituent that has the maximum F_0 in that constituent. The program also displays "SENTB" at sentence boundaries, although none are shown in the example of Figure 3.

The positions of assigned syntactic boundaries and maximum F_0 in the constituent can also be displayed on the interactive graphical display, for on-line analysis of boundary detection results (cf. Lea, Medress, and Skinner, 1972a, p. 26).

Several improvements could be made in the syntactic predictions and the detailed workings of Lea's boundary detection algorithm, but, for the most part, they have not been incorporated into the present studies. One improvement that has been used in the syntactic predictions of boundaries is that boundaries are not predicted between pronouns and following verbals. Also, in the hand analysis of the Rainbow Script, to be described in section 3.2 below, one refinement in the algorithm was incorporated which eliminated false boundaries resulting from variations in F_0 that only last for one 10-millisecond time sample.

The boundary detection results take on new significance when they are used in the algorithm for stressed syllable location to be described in section 3.4. Not only do the detected boundaries provide a means of segmenting continuous speech into syntactically relevant units, but they provide critical data used in the procedure for locating stressed syllables.

Certain refinements in the boundary detector, and the total algorithm for stressed syllable location, are to be implemented on the research facility as will be discussed in section 4.

3. EXPERIMENTS ON SYNTACTIC BOUNDARIES AND STRESS PATTERNS

3.1 Experimental Design

The strategy for speech recognition outlined in section 1 suggests the need for methods of demarcating constituents, finding stressed syllables, and doing a partial distinctive features analysis on the reliable data within the stressed syllables. A method for demarcating constituents has been developed (Lea, 1972b, 1973a), but its implementation and refinement at Univac had to be tested with extensive speech data. Methods for finding stressed syllables from acoustic data had to be developed. These will provide critical guidelines to procedures for partial distinctive features estimation and syntactic parsing.

In an earlier report (Lea, Medress, and Skinner, 1972a) a three-fold experimental effort in stressed syllable analysis was suggested. This involved: (1) obtaining linguistic predictions of stress patterns by applications of syntactic analysis and appropriate stress rules and vowel reduction rules; (2) determining listeners' actual perceptions of stressed, unstressed, and reduced syllables; and (3) locating stressed syllables by analysis of F_0 contours and intensity contours obtained from the acoustic data. All this study of predicted, perceived, and acoustically-determined stress patterns was to be conducted on readings of the "Rainbow Script", a semantically-connected text of six declarative sentences. Detailed advantages of both this text and the overall experimental design were presented in the earlier report (Lea, Medress, and Skinner, 1972a, pp. 32-40).

The experiments reported on here are a modification of the suggested experiments. The linguistic predictions of stress patterns are not presented here. It is a major task to select adequate stress rules from the many conflicting ones in the literature (Chomsky and Halle, 1968; Halle and Keyser, 1971; Vanderslice and Ladefoged, 1971; Bresnan, 1971; 1972; Lakoff, 1972; Berman and Smazosi, 1972) and to apply them rigorously to arbitrary selected texts. Hopefully, such studies will be pursued later.

On the other hand, the original experiments on stress patterns in the "Rainbow Script" have been extensively expanded, to include studies of detected syntactic boundaries in the script when read by six talkers, and to include studies of stress patterns and syntactic boundaries in other speech texts. A major breakthrough has

been accomplished in the development of an algorithm for locating stressed syllables from F_0 contours and energy contours.

These experiments on syntactic boundaries and stress patterns in several spoken texts have been described in detail in a recent report (Lea, 1973h). In this section (section 3), we will briefly summarize the results of these extensive experiments. Section 3.2 presents the speech texts selected for study. The success in detecting syntactic boundaries is discussed in section 3.3. Stress patterns in the texts, as perceived by a panel of listeners on several trials, are summarized in section 3.4. The algorithm for locating stressed syllables, and its success in locating most stressed syllables in the selected texts, are presented in section 3.5.

3.2 Speech Texts Selected for Analysis

To test the algorithms for boundary detection and stressed syllable location, speech texts had to be chosen, recorded, submitted to listeners for stress perceptions, and analyzed by the computer programs. The primary text chosen for these studies was the first paragraph of the "Rainbow Passage" (Fairbanks, 1940). It reads as follows:

> "When the sublight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow."

This text (hereinafter called the Rainbow Script) has been used extensively in studies of prosodic patterns in speech, and has the advantage of being a well-known semantically-connected text of declarative sentences, with a variety of grammatical phrase structures (cf. Lea, Medress, and Skinner, 1972a). It was recorded by six talkers (four male, two female) in a quiet room at Purdue University.

In texts like the Rainbow Script, the factors determining positions of stress within words (lexical stress) are compounded with sentence structure effects on stress (cf. Chomsky and Halle, 1968; Halle and Kayser, 1971). Another text which was composed of only monosyllabic words was also analyzed, to eliminate or minimize lexical

stress effects. This text, read by two of the six talkers who had read the Rainbow Script, is the first paragraph of a short story:

> "John and I went up to the farm in June. The sun shone all day, and wind waved the grass in wide fields that ran by the road. Most birds had left on their trek south, but old friends were there to greet us. Piles of wood had been stacked by the door, left there by the man who lives twelve miles down the road. The stove would not last till dawn on what he had cut, so I went and chopped more till the sun set."

Lea (1972a, b) had previously processed recordings of this text (hereinafter referred to as the Monosyllabic Script) for constituent boundary detection at Purdue University. Comparing his previous results with the boundary detections found by the Univac implementation of his algorithm helped verify the new algorithm.

Both the Rainbow Script and the Monosyllable Script involve read speech, all of declarative structure. To evaluate the boundary detection and stressed syllable location techniques with questions, commands, and declaratives of direct utility in man-machine interactions, thirteen sentences were selected from actual recordings by five contractors who are developing speech understanding systems for the Advanced Research Projects Agency. Most of these sentences were <u>not</u> read, but were composed on the spot in simulated protocols of man-machine interaction. The semantic context of each sentence was pertinent to a particular task domain adopted by the builder of a speech understanding system, such as retrieving information on lunar rock samples (Woods, 1971), other information-retrieval tasks, instructing a robot to move objects in a block world (Walker, 1973), or voice programming.

These thirteen sentences are as follows:[1]

1. (LS21) Who's the owner of utterance eight?
2. (LM13) Display the phonemic labels above the spectrogram.

---

1. The first letter of the code identifying each sentence, as shown within the parentheses of this list, indicates the ARPA contractor which recorded the sentence (B = Bolt, Beranek, and Newman; C = Carnegie Mellon University; D = Systems Development Corporation; L = Lincoln Laboratories; and R = Stanford Research Institute). The second letter, when it appears, identifies which talker from that organization spoke the associated sentence, or, in the case of CV codes, it marks the task as voice programming. Numbers in the code indicate the order in which the sentence appeared in that organization's protocol of utterances. This complex code is included here since these same utterances are being studied, under such identifiers, by various ARPA contractors.

3. (B27) Do any samples contain troilite?
4. (B10) What is the average uranium lead ratio for the lunar samples?
5. (RB6) Do you have any right square boxes left?
6. (RB16) Put the other red block on the red block.
7. (LM3) Who is the owner of utterance eight?
8. (B35) Do any samples contain tridymite?
9. (RA19) Would you move the stack of right circular cylinders to the right by half a square?
10. (RC8) Place the red triangle two squares back from the front of the floor in the middle.
11. (CV1300) Alpha becomes alpha minus beta.
12. (CV2300) Alpha gets alpha minus beta.
13. (D10) Repeat where key work equals Gauss elimination or key word equals eigenvalues.

The recordings of all these speech texts provide a total of 379 predicted syntactic boundaries and 1128 syllables for evaluating the effectiveness of the boundary detection and stressed syllable location algorithm.

3.3 Syntactic Boundaries Detected in the Selected Texts

The detailed results of boundary detection for the six talkers reading the Rainbow Script, the two talkers reading the Monosyllabic Script, and the one recording of the ARPA Sentences (involving ten talkers) are presented in a recent report (Lea, 1973b, section 3). These results are summarized in Table I. The percentages of all predicted boundaries that were detected for each text are shown in the second column (with talkers pooled as shown in the leftmost column). The spontaneous ARPA Sentences show the lowest overall detection score, partly due to the monotonic intonation with which some of them were spoken, and partly due to unusual F_0 inflections and hesitation pauses in these simulations of man-machine interactions. However, the score of 74% is not drastically different from the average score of 81% obtained in previous studies of written texts and suggests that most boundaries can be detected in utterances suitable for man-machine interactions.

Also shown in Table I (the third column) are the total numbers of correctly detected boundaries that had been predicted, the numbers of "extra" detected boundaries that are apparently due to breaks between minor syntactic constituents (fourth column),

TABLE I.

SUMMARY OF
BOUNDARY DETECTION SCORES

Text	Percent Predicted Boundaries Detected	Number of Predicted Boundaries that Were Detected	Number of Extra Detected Boundaries	Number of False Detected Boundaries
Rainbow Script: 6 Sentences 6 Talkers 252 Boundaries	79%	198	29	9
Monosyllabic Script: 5 Sentences 2 Talkers 84 Boundaries	83%	70	12	10
ARPA Sentences: 13 Sentences Mixed Talkers 50 Boundaries	74%	37	9	7+8 pauses

and the numbers of "false" detected boundaries which don't appear to be related to any syntactic boundary (fifth column). The Rainbow Script was analyzed by a hand analysis strictly following Lea's algorithm, but incorporating one refinement that eliminated false boundaries resulting from sudden changes in F_0 (at boundaries between consonants and vowels) that were not sustained for 20 milliseconds or more. The other texts were analyzed by the Univac implementation of Lea's original algorithm, without this refinement. The percentages of all detected boundaries that were false were thus proportionately higher for the results from the computer algorithm. The ARPA Sentences also included a number of extraneous hesitation pauses, which were falsely characterized as clause boundaries by the boundary detector. These are listed separately in the last column of Table 1.

3.4 Perceived Stress Patterns in the Selected Texts

Listeners' perceptions of stress provide a standard by which stress detections from acoustic cues can be tested. Previous studies have attempted to determine how listeners' judgments of stress vary as certain acoustic features are varied, usually in synthesized speech (cf. Lea, Medress and Skinner, 1972a, pp. 32-40). However, few such studies have been concerned with the stress patterns throughout sentences; most work was done on isolated words such as minimal pairs of noun versus verb (permit/permit, etc.). Some attempts have been made to determine listeners' perceptions of the most stressed syllable in a sentence, or which of two specific syllables is more stressed, or whether a specific single syllable is or is not stressed. The present experiments extend studies to all syllables in the sentences.

Three listeners (WAL, MFM, and TES) each individually heard (through earphones) the Rainbow Script as recorded by the six talkers, the Monosyllabic Script as recorded by the two talkers, and the ARPA Sentences. Each listener heard clauses or sentences, or other extended portions of the text, repeated at will, by the listener's rewinding and replaying of the tape. The listener was instructed to mark (in whatever way he chose), for each syllable, whether he heard that syllable as stressed, unstressed, or reduced. To facilitate marking for each syllable, each script was typed on a sheet of paper with vertical slashes between syllables (except for the Monosyllabic Script, in which each word is one syllable). A mark was required for each syllable (between two slash marks). The listener completed one such sheet for each repetition and each talker and text.

The Rainbow Script was specifically separated into clauses separated by long pauses, to aid the rewinding and replay, while the other recordings were not. The listeners endeavored to rewind far enough to always hear an entire clause, to have a constant context within which to judge relative stress levels. Each listener could listen to the tape portions as often as necessary to mark each syllable. He was free to back up the tape at his choice, and no time limit or procedural constraints were placed on him.

Each listener repeated the perception test three times (with no less than three days between trials) to establish listener consistency from one time to another. Also, to establish that the actual speech heard was playing a role in stress judgments, the listeners were also asked to report their stress judgments given only the written text. This test with no speech was included to determine whether the listener's presuppositions, internal "theory" of expected stress patterns, or own way of speaking the sentences was the sole source of his decisions or whether the scoustic data actually was supplying cues to stress patterns. These no-speech stress judgments were also obtained in three repetitions, spaced three or more days apart, to test their repeatability.

The Rainbow Script contains 128 syllables, the Monosyllabic Script 87 syllables, and the ARPA Sentences 171 syllables. With three repetitions with speech, three without speech, three listeners, and with the various speakers involved, this totals about 28,000 judgments of stress levels for syllables in connected texts. The detailed analysis of these extensive results was presented in a recent report (Lea, 1973b, section 4). Here we shall summarize the major conclusions from such studies:

1. Different listeners assign different stress levels to the same syllables, presumably based on how they individually define the boundaries between categories of stressed, unstressed, and reduced syllables. Their confusions are not seriously increased or decreased in going from individual talker to talker, or from text to text (except when questions are introduced; see point 8 below).

2. Listeners WAL and MFM, who have been shown by previous experiments to yield stress perceptions very much like those of other listeners used previously, differed in as much as 25 to 30% of their majority decisions about stress levels

(compiled from three trials). However, only about 5% of all syllables were confused between the categories stressed and unstressed. Thus, judgments of which syllables were stressed agreed very well between listener WAL and listener MFM.

3. Listener TES differed from the other two listeners on about half of his stress decisions. About 20 to 25% of all syllables were labelled stressed by other listeners, but unstressed by TES. He actually even labeled as reduced some syllables labelled stressed by the other listeners. Also, listener TES labelled substantial percentages (as much as 15%) of all syllables as stressed on one trial and unstressed on another. Future studies should incorporate a procedure for rejecting such listeners who provide inconsistent judgments about stressed syllables.

4. From repetition to repetition of the perception tests, listeners WAL and MFM individually showed quite stable judgments as to which syllables were stressed. An average of 5% of all syllables were confused between stressed on one trial and unstressed on another trial. They thus provide a reasonably stable "standard" as to which syllables are stressed, for comparison with algorithm results.

5. Majority votes obtained from 3 or more trials should be used to partially obliterate the 5% deviations in assignment of stressed syllables from trial to trial. No stressed syllable location algorithm need find more than 95% of all syllables perceived as stressed, since it can hardly be more "accurate" than one perception trial is in predicting the perceptions to be attained on another trial.

6. Since listeners agreed in many of the differences they assigned to the stress patterns of different talkers reading the same text, the acoustic data appears to play at least some role in stress perceptions. However, since listener-to-listener confusions and most repetition-to-repetition confusions were not significantly increased when only the written text was used, it appears that the listeners also make use of a reasonable stable internal theory for stress assignment.

7. When listeners had not done the perception tests with speech before they did the stress assignments from the written text alone (as with the Monosyllabic

Script and the ARPA Sentences), their majority judgments without speech differed more from their majority perceptions with speech than the repetition-to-repetition with speech (or without speech) had differed. Thus, while stress judgments without speech are as consistent from listener-to-listener and from repetition-to-repetition as are perceptions with speech, the judgments made without speech are significantly different from those made with speech. In particular, perceived stress patterns for spontaneous utterances are not reliably obtained from judgments based only on the written text.

8. Questions (especially yes-no questions) appear to yield more confusions in stress levels (from repetition-to-repetition) than other sentence structures (declaratives or commands), and show greater variability from listener-to-listener.

In summary, the stress perceptions obtained from the trials with speech, by using majority decisions for each listener, and pooling results for the listeners, provide a "standard" of stress assignment which is stable to within about 5%. This permits comparisons (to within 5%) between perceived stressed syllables and stressed syllables located by algorithm from the acoustic data.

3.5 Stressed Syllable Location from the Acoustic Data

Based on the extensive previous studies of acoustic correlates of stress (see review by Lea, 1972b, sections 5.1 and 5.2), an algorithm for locating stressed syllables has been developed. The algorithm assumes that local increases in F_0 and high energy integral are the most reliable correlates of stressed syllables. The increasing F_0 near the beginning of each constituent detected by the boundary detector is assumed to be attributable to the first stressed syllable in the constituent (Lea, 1973b, section 5). A stressed "HEAD" to the constituent is thus associated with a portion of the speech which is high in energy with rising F_0, and bounded by substantial (5 dB or more) dips in energy. Other stressed syllables in the constituent are expected to be accompanied by local increases in F_0. Since the usual ("archetype") shape of the F_0 contour in a constituent is a rapid rise followed by a gradual fall in F_0, we expect that local 'increases' in F_0 due to later stressed syllables will show local rises above the gradually falling F_0 contour, even if F_0 does not rise absolutely near the stressed

syllable. The stressed syllable is located within a high-energy-integral region near this local rise above the archetype F_0 contour.

Lea (1973b, section 5) presented a detailed description of an algorithm for locating stressed syllables, based on the strategy just outlined, and reported the results of a hand analysis of the selected texts following the algorithm. Table II shows the overall comparison between the algorithmically located stressed syllables and the listeners' perceptions of stressed syllables. For each text and talker (leftmost column), the table gives,in the second column, the number of syllables perceived as stressed by two or more listeners (with at most one listener saying the syllable was unstressed) and the number of those syllables (perceived as stressed) that the algorithm correctly located within high energy portions called "stressed syllables". Occasionally a stretch of speech was located by the algorithm that did not enclose any syllable perceived as stressed by the listeners. This gave the numbers of "false" locations shown in the next to the rightmost column. Dividing the number of false locations by the total number of algorithmically located portions gives the percent of all locations that were false (rightmost column).

While scores varied somewhat from text to text and talker to talker, the overall scores of 78% to 98% (average, 85%) correct location of stressed syllables are very encouraging. The Monosyllabic Script, with its fewer reduced syllables and more prominant stresses on monosyllabic content words, yielded quite high scores. The spontaneous ARPA Sentences, which were more monotone and which gave some difficulties to the boundary detection algorithm, showed the lowest stressed syllable location scores.

The false alarm rates were fairly high, ranging from 7% up to 28%. Some of the false alarms will be eliminated by improvements in the boundary detector. Some other 'false' locations are not necessarily bad, since one or two listeners did perceive those syllables as stressed. A few of the false alarms may be eliminated by not demanding stressed HEADS in short constituents (such as those less than 200 ms in duration). Further studies are needed to reduce false alarm rates and simultaneously maintain or improve the scores for correct locations.

Ultimately, the design of a better algorithm for stressed syllable location must be based on a strategic decision as to whether it is better to have some false alarms and correspondingly increase the success in correct location or to have little or no

TABLE II

STRESSED SYLLABLE LOCATION SCORES

Text	Number of Stressed Syllables Perceived by Two or More Listeners (P)	Number of Such Syllables Correctly Detected (D)	Percent Stressed Syllables Correctly Detected (D/P x 100%)	Number of 'False' Locations (F)	Percent of All Locations That Were 'False' (F/D+ F) x 100%
RAINBOW					
ASH	51	43	84%	3	7%
GWH	45	44	98%	7	14%
WB	47	38	81%	15	28%
JP	48	42	88%	15	26%
PB	50	39	78%	6	13%
ER	49	43	88%	3	7%
MONOSYLLABIC					
ASH	41	37	90%	8	18%
GWH	41	39	95%	14	26%
13 ARPA SENTENCES	70	56	80%	14	20%

false alarms but at the sacrifice of lower scores in correct location. This will substantially depend upon the specific use of stressed syllable information in other aspects of the speech understanding system. For guiding distinctive features estimation procedures, all that might come from having a few false locations is that distinctive features analysis may occasionally be applied (perhaps wastefully or with some difficulty) in the somewhat-less-reliably-encoded unstressed syllables.

4. CONCLUSIONS AND FURTHER STUDIES

Substantial progress has been made on the development of analysis tools and algorithms that are critical to providing prosodic aids to speech recognition. The method for determining fundamental frequency from autocorrelation of the center-clipped time waveform has proven very reliable and accurate in the analysis of over 400 seconds of speech. Formant tracking from peak picking on smoothed spectra provided by linear prediction has been implemented and tested, and enhanced by incorporating the successful refinements of other ARPA researchers. The constituent boundary detection program has been implemented on the Univac speech research facility, and tested with extensive data. When coupled with the algorithm for stressed syllable location (which is not yet implemented as a computer program), these tools will provide significant aspects of the basic strategy of locating reliable, clearly encoded data for speech recognition. The connection to the ARPANET, when completed, will provide means for integrating such prosodic and distinctive features information with other aspects of speech understanding systems.

In this report, methods have been described for segmenting speech into grammatical phrases and identifying stressed syllables in continuous speech. The program for detecting syntactic boundaries from fall-rise patterns in voice fundamental frequency contours has been shown, both by the present study and by previous studies, to succeed in finding over 80% of all syntactically predicted boundaries between major syntactic units. It also, however, detects some syntactic boundaries not predicted by the intuitive constituent structure analysis previously applied, and detects false boundaries not apparently related to syntactic structure, such as at consonant-vowel boundaries.

The algorithm for stressed syllable location has succeeded in locating about 85% of all syllables perceived as stressed by the majority votes of a panel of listeners. The procedure identifies stressed syllables with high-energy-integral portions of the speech which exhibit rising or non-falling F_0, but it does so in a way which makes use of constituent boundaries and archetype F_0 contours. Simpler procedures might conceivably work as well, and there is obviously room for improvement in the present location scores.

Besides such algorithmic results, the other major aspect of research reported herein has been concerned with the perceptions of stress levels by three listeners. Two listeners were found to agree in their perceived stress levels for most of the individual syllables in the Rainbow Script and Monosyllabic Script, and ARPA man-machine interaction sentences. They differed on only about 5% of all syllables as to whether they were stressed or not, and each of them showed only about 5% confusions in decisions about stressed syllables from one trial to another. Unstressed and reduced levels were much more frequently confused. A third listener differed from the other two listeners on about half of his stress level judgments. About 20 to 25% of all syllables were labelled stressed by the other listeners, but unstressed by this third listener. This listener also labelled substantial percentages of all syllables as stressed on one trial and unstressed on another. Such listeners who are inconsistent in their own judgments and who differ dramatically from other listeners should be excluded in any attempts to establish standards about which are the actual "stressed syllables" in connected speech.

The listeners appear to be as consistent in their assignments of stress levels given only the written text as they are in their assignments when listening to the speech recordings. However, their judgments without speech do not correspond well with their judgments with speech if the speech is spontaneous (that is, not produced by speakers reading written texts). Listeners apparently differ most dramatically from each other, and yield more confusions in stress levels from repetition to repetition, when yes-no questions are involved.

The majority stress perceptions from three trials by each listener, when pooled for all listeners, provide a "standard" for determining all stressed syllables which is stable to within about 5%. This is suitable for evaluating an algorithm for locating stressed syllables to within a 5% tolerance in overall location scores.

Several forms of further work are needed. The program for constituent boundary detection can be refined to produce fewer false alarms by requiring each new F_0 maximum or minimum to remain beyond the 7% thresholds for at least 20 ms. It would be desirable to remove or decrease the strict dependence on a fixed (7%) threshold for F_0 changes, and to incorporate an overall confidence measure for each boundary, based on the percentage decrease in F_0 before the apparent boundary, the percentage

4. CONCLUSIONS AND FURTHER STUDIES

Substantial progress has been made on the development of analysis tools and algorithms that are critical to providing prosodic aids to speech recognition. The method for determining fundamental frequency from autocorrelation of the center-clipped time waveform has proven very reliable and accurate in the analysis of over 400 seconds of speech. Formant tracking from peak picking on smoothed spectra provided by linear prediction has been implemented and tested, and enhanced by incorporating the successful refinements of other ARPA researchers. The constituent boundary detection program has been implemented on the Univac speech research facility, and tested with extensive data. When coupled with the algorithm for stressed syllable location (which is not yet implemented as a computer program), these tools will provide significant aspects of the basic strategy of locating reliable, clearly encoded data for speech recognition. The connection to the ARPANET, when completed, will provide means for integrating such prosodic and distinctive features information with other aspects of speech understanding systems.

In this report, methods have been described for segmenting speech into grammatical phrases and identifying stressed syllables in continuous speech. The program for detecting syntactic boundaries from fall-rise patterns in voice fundamental frequency contours has been shown, both by the present study and by previous studies, to succeed in finding over 80% of all syntactically predicted boundaries between major syntactic units. It also, however, detects some syntactic boundaries not predicted by the intuitive constituent structure analysis previously applied, and detects false boundaries not apparently related to syntactic structure, such as at consonant-vowel boundaries.

The algorithm for stressed syllable location has succeeded in locating about 85% of all syllables perceived as stressed by the majority votes of a panel of listeners. The procedure identifies stressed syllables with high-energy-integral portions of the speech which exhibit rising or non-falling F_0, but it does so in a way which makes use of constituent boundaries and archetype F_0 contours. Simpler procedures might conceivably work as well, and there is obviously room for improvement in the present location scores.

success by looking for all F_0 rises or upward inflections and choosing the high energy portion nearest such places, without use of boundaries and archetype contours in his procedures?

More extensive experiments are needed wherein the various variables of sentence type, talker, lexical forms, phonetic content, position in sentence and intonation contour, and such could be independently controlled. Texts for such studies are now being designed (cf. Lea, Medress, and Skinner, 1972a, pp. 56-57), and such studies will be conducted. In particular, such studies can test further the apparent difficulty in listeners' assignments of stress within yes-no questions, and the relative successes in boundary detection and stressed syllable location within questions versus declaratives or commands.

The application of boundary detections and stressed syllable locations to guiding a partial distinctive features analysis must yet be done. Until such details of the distinctive features analysis are better defined, the question cannot be resolved as to whether higher "hit" rates or lower "false alarm" rates are more important to attain in the boundary detection or stressed syllable location algorithm. Also, techniques must be explored for applying boundary and stressed syllable information to the aid of syntactic parsers. Such efforts will be critical to implementing the proposed speech recognition strategy at Univac.

Some thought has been given to the possibility of speeding up the present analysis procedures and improving the potential for eventually attaining approximately real-time analysis. One aspect which has been explored briefly is the possibility of using absolute addition as an alternative to multiplication in forming the autocorrelation function used in fundamental frequency analysis. That is, for the AUTOCORRELATION EQUATION given in the Appendix, substitute this equation:

$$A_j = \sum_{i=1}^{N} \left| C_i + Z_{i+j-1} \right| \qquad j = O_L, O_L+1, O_L+2, \ldots, O_M$$

This formulation has the advantage of savings in computation time (addition as opposed to multiplication), and is simpler to potentially implement in real-time hardware. However, the savings in computation time for the software algorithm (approximately 10%) is not appreciable, since both formulations bypass computations during zero elements of the center-clipped vector.

The absolute addition formulation was used in the computation of fundamental frequency on some of the ARPA test sentences, with results comparable to those achieved when using the multiplication formulation. Alternative methods of rapidly obtaining appropriate energy measures must also be considered.

In summary, this research has yielded substantial success in constituent boundary detection and stressed syllable location, using a computer program for boundary detection and a general strategy for stressed syllable location. An adequate method has been devised for obtaining listeners' judgments of which syllables in connected speech are stressed, unstressed, or reduced. Further controlled studies are needed to refine boundary predictions and detections; to predict stress patterns linguistically; to implement the algorithm for stressed syllable location; to improve the basic methods of tracking fundamental frequency, energy, and formants; to test these prosodic algorithms with designed speech texts; and to relate these prosodic aids to processes of distinctive features estimation and syntactic parsing.

5. REFERENCES

BERMAN, A., and SZAMOSI, M. (1972), Observations on Sentential Stress, Language, vol. 48, 304-325.

BRESNAN, J. (1971), Sentence Stress and Syntactic Transformations, Language, vol. 47, pp. 157-81.

BRESNAN, J. (1972), Stress and Syntax: A Reply, Language, vol. 48, pp, 326-42.

CHOMSKY, N. and HALLE, M. (1968), The Sound Pattern of English. New York: Harper and Row.

FAIRBANKS, G. (1940), Voice and Articulation Drillbook. New York: Harper and Row.

HALLE, M. and KEYSER, S. J. (1971), English Stress. New York: Harper and Row.

LAKOFF, G. (1972), The Global Nature of the Nuclear Stress Rule, Language, vol. 48, 285-303.

LEA, W. A. (1972a), An Approach to Syntactic Recognition without Phonemics. Proc. 1972 Intern. Conf. on Speech Commun. and Processing. Newton, Mass.: pp. 198-201. A revised version of this paper has been accepted for publication in the IEEE Transactions on Audio and Electroacoustics.

LEA, W. A. (1972b), Intonational Cues to the Constituent Structure and Phonemics of Spoken English, Ph.D. Thesis, School of E. E., Purdue University.

LEA, W.A. (1973a), Segmental and Suprasegmental Influences on Fundamental Frequency Contours. Presented at the Symposium on Consonant Types and Tone, University of Southern California, Los Angeles, March 9-10, 1973. To appear in Consonant Types and Tone (Proceedings of the First Annual Southern California Round Table in Linguistics), University of Southern California.

LEA, W.A. (1973b), Syntactic Boundaries and Stress Patterns in Spoken English Texts, Univac Report No. PX 10146, Univac Park, St. Paul, Minnesota.

LEA, W.A., MEDRESS, M.F., and SKINNER, T.E. (1972a), Prosodic Aids to Speech Recognition I: Basic Algorithms and Stress Studies, Univac Report No. PX 7940, Univac Park, St. Paul, Minnesota.

LEA, W.A., MEDRESS, M.F., and SKINNER, T.E. (1972b), Use of Syntactic Segmentation and Stressed Syllable Location in Phonemic Recognition. Presented at the 84th Meeting, Acoustical Society of America, Miami Beach, Florida, Nov. 27-30, 1972.

MAKHOUL, J. and WOLF, J. (1972), Linear Prediction and the Spectral Analysis of Speech, Technical Report, ARPA Contract No. DAHC-71-C-0088, Bolt Bernanek and Newman Report No. 2304.

SONDHI, M. M. (1968), New Methods of Pitch Extraction, IEEE Trans. on Audio and Electroacoustics, vol. AU-16, pp. 262-266.

VANDERSLICE, R, and LADEFOGED, P. (1971), Binary Suprasegmental Features. Working Papers in Phonetics No. 17, Phonetics Lab., Univ. of Calif. at Los Angeles, pp. 6-23.

WALKER. D, E. (1973), Speech Understanding Research, Annual Technical Report prepared for ARPA, Contract DAHC04-72-C-0009, Stanford Research Institute, Menlo Park, California.

WOODS, W. A. (1971), The Lunar Sciences Natural Language Information System, BBN Report No. 2265, NASA Contract NAS9-1115, Bolt Beranek and Newman, Cambridge, Massachusetts.

APPENDIX: AUTOCORRELATION METHOD FOR DETERMINING FUNDAMENTAL FREQUENCY

The following is a method for determining the fundamental frequency of a speech signal using autocorrelation.

A digitized time vector T is selected for processing. The dimension N of this vector is a function of the sampling rate of the analog signal, such that the time segment contains at least two periods of the minimum fundamental frequency to be detected (see Figure A1).

$$T: \quad t_1, t_2, t_3, \ldots, t_N$$

The signal energy (IENERG) in the time segment corresponding to vector T is ten times the logarithm of the autocorrelation of T at zero offset.

$$IENERG = 10 \cdot LOG \left(\sum_{i=1}^{N} t_i^2 \right)$$

To avoid potential integer overflow in the energy computation, IENERG is computed as:

$$IENERG = 10 \cdot LOG \left(\sum_{i=1}^{N} (t_i/10)^2 \right) + 20$$

which serves as an approximation since

$$10 \cdot LOG \left(\sum_{i=1}^{N} (t_i/10)^2 \right) \approx 10 \cdot LOG \left(\sum_{i=1}^{N} t_i^2 \right) - 20$$

If the signal energy does not exceed a threshold (IETHRS), the time segment is declared to be unvoiced and thus the fundamental frequency is recorded as zero.

$$IENERG < IETHRS, \text{ implies } F_0 = 0$$

Provided the signal has enough energy to be considered for further fundamental frequency processing, the time vector T is next center clipped at a percentage ICUT of the absolute maximum t_M in the time segment. This technique of center clipping the signal prior to autocorrelation is chiefly attributed to Man Sondhi of Bell Telephone Laboratories ("New Methods of Pitch Extraction," IEEE, June 1968).

$$t_M = \text{MAX} \left| (t_1, t_2, t_3, \ldots, t_N) \right|$$

$$C_i = \begin{cases} 0 & \text{if } |t_i| < \text{ICUT} \cdot t_M/100 \\ t_i - \text{ICUT} \cdot t_M/100 & \text{if } t_i \geq \text{ICUT} \cdot t_M/100 \\ t_i + \text{ICUT} \cdot t_M/100 & \text{if } t_i \leq -\text{ICUT} \cdot t_M/100 \end{cases}$$

$$C: \; C_1, C_2, C_3, \ldots, C_N$$

The number of changes of state (IPZERO) from zero to non-zero and from non-zero to zero elements in the center clipped vector C must not equal or exceed a threshold (IPTHRS) or the time segment is declared unvoiced. (In voiced speech, zero or non-zero elements of the center clipped time vector will occur in contiguous bands.)

$$\text{IPZERO} \geq \text{IPTHRS}, \text{ implies } F_0 = 0$$

If the time segment contains sufficient energy to be classified as a speech signal and if the center clipped vector has a sufficient number of contiguous zero elements to further classify the time segment as a potential voiced speech signal, a band of fundamental frequency detection is then defined such that F_L = minimum F_0 and F_M = maximum F_0 to be detected.

For a sampling rate of R KHz, the time in msec between successive samples is DT: $DT = R^{-1}$.

From F_L and F_M (the F_0 frequency limits), are defined O_L and O_M, the autocorrelation time offset limits.

$$O_L = \frac{1000}{(F_M \cdot DT)} + 1$$

$$O_M = \frac{1000}{(F_L \cdot DT)} + 1$$

The vector C is circularly autocorrelated to define a new vector A.

$$Z: \; C_1, C_2, C_3, \ldots, C_N, C_1, C_2, C_3, \ldots, C_N$$

$$A_j = \sum_{i=1}^{N} C_i \cdot Z_{i+j-1} \qquad j = O_L, O_L + 1, O_L + 2, \ldots, O_M$$

AUTOCORRELATION EQUATION

The maximum value of the autocorrelation vector and its associated offset index over the bounds defined by O_L and O_M are determined.

$$V = \text{MAX } (A_{O_L}, A_{O_L+1}, A_{O_L+2}, \ldots, A_{O_M})$$

F = offset index at which maximum occurs
(subscript of A)

If V does not exceed a percentage IVTHRS of the autocorrelation function at zero offset, then the time segment is declared unvoiced.

$$V \leq \text{IVTHRS} \cdot A_1/100, \text{ implies } F_0 = 0$$

If V does exceed IVTHRS • $A_1/100$, then the value of A at F/2 is examined to determine if the value at this offset also exceeds the threshold IVTHRS • $A_1/100$. If in fact it does, F is redefined as F/2.

$$V > \text{IVTHRS} \cdot A_1/100 \text{ and } A_{F/2} > \text{IVTHRS} \cdot A_1/100$$

then F = F/2

This process is repeated iteratively until $F/2 < O_L$ or until a maximum of four octaves have been investigated.

Now, if the process has surpassed the energy threshold, the number of zero elements in the center clipped vector threshold, and the autocorrelation offset percentage of A_1 threshold, then F is defined as non-zero and is computed as follows:

$$F_0 = \frac{1}{(F-1) \cdot DT/1000}$$

Figure A2 graphs this relationship, and shows that the resolution of the F_0 estimate decreases with increasing F_0.

Parameter values which have yielded successful per time segment fundamental frequency determination results for a 10 KHz sampling rate are: N = 512, ICUT = 30, IPTHRS = 150, IVTHRS = 28, IETHRS = 65.

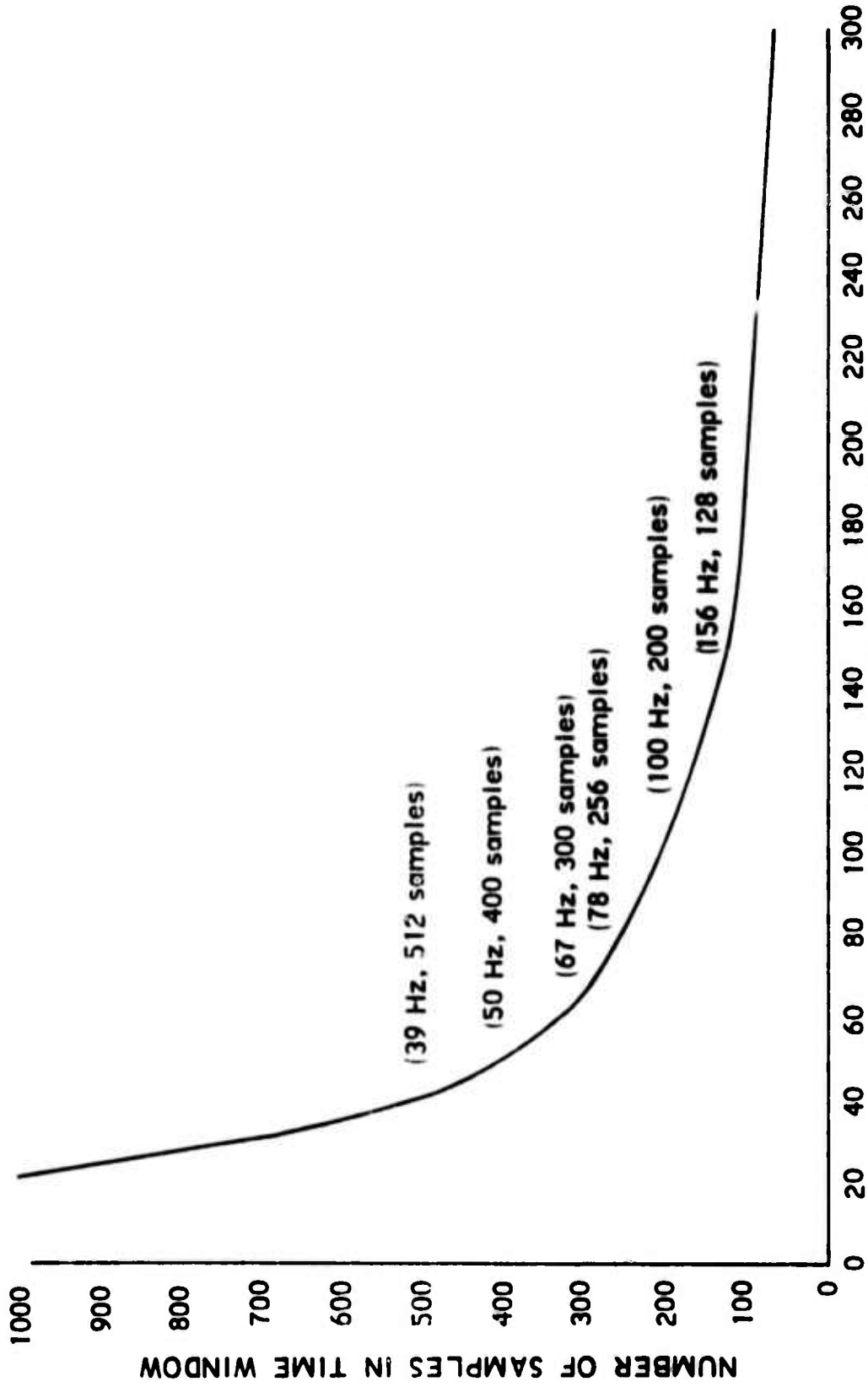

Figure A1. Time Window Width for F_0 Analysis (10 KHz Sampling Rate)

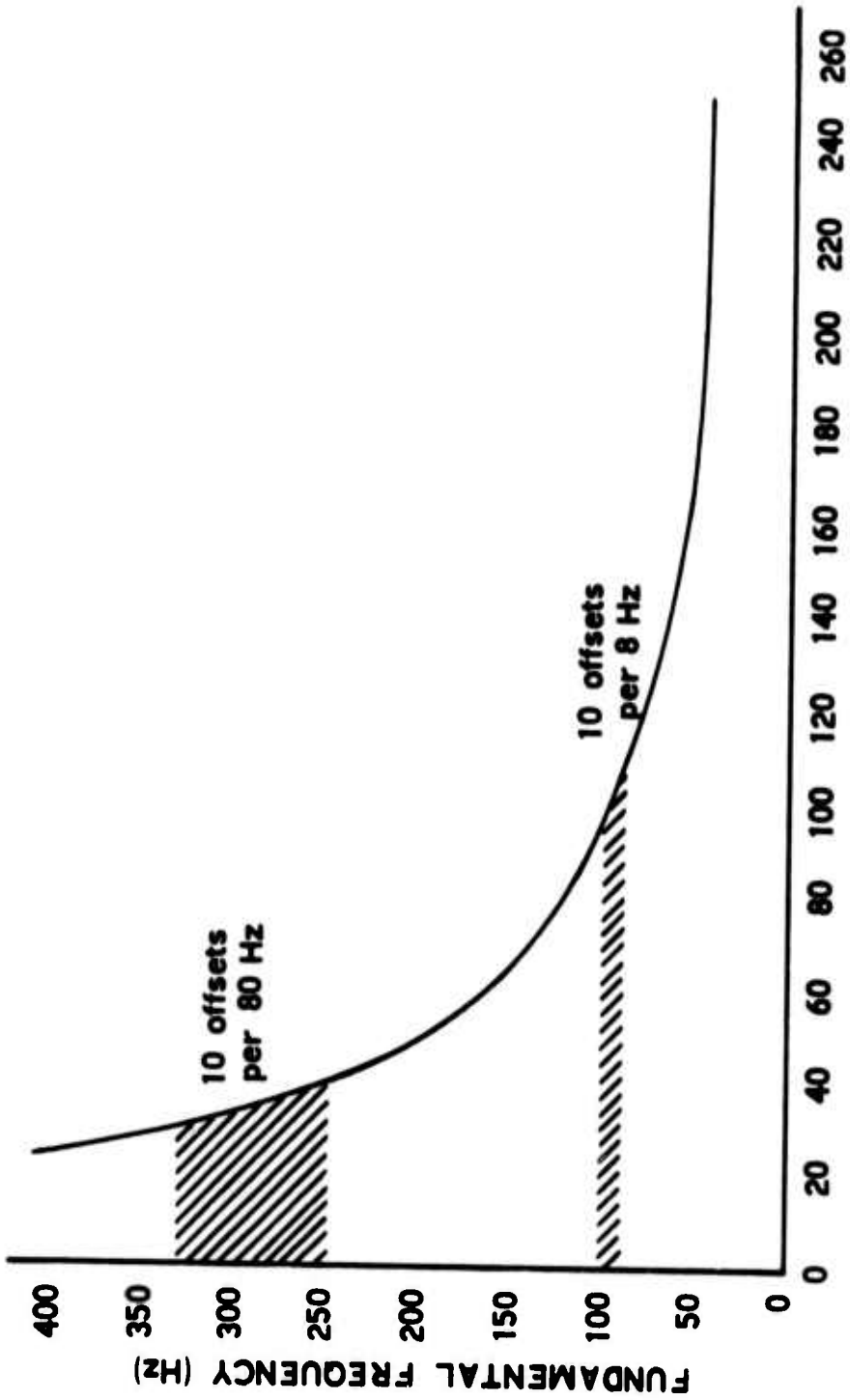

Figure A2. Relationship between F_0 and Autocorrelation Vector Offset (10 KHz Sampling Rate)